

Combinatorics of Genome Rearrangements

Guillaume Fertin, Anthony Labarre, Irena Rusu, Éric Tannier and Stéphane Vialette

The MIT Press
Cambridge, Massachusetts
London, England

© 2009 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

For information about special quantity discounts, please email special_sales@mitpress.mit.edu

This book was set in Times New Roman and Syntax on 3B2 by Asco Typesetters, Hong Kong.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Combinatorics of genome rearrangements / Guillaume Fertin . . . [et al.].

p. cm. — (Computational molecular biology)

Includes bibliographical references and index.

ISBN 978-0-262-06282-4 (hardcover : alk. paper) 1. Translocation (Genetics)—Mathematical models.
2. Translocation (Genetics)—Data processing. 3. Combinatorial analysis. 4. Genomics—Mathematics.

I. Fertin, Guillaume, 1972– II. Series.

[DNLM: 1. Gene Rearrangement. 2. Genome. 3. Models, Genetic. QU 470 C731 2009]

QH462.T7C66 2009

572.8'77—dc22

2008042152

10 9 8 7 6 5 4 3 2 1

Contents

Preface	xiii
Acknowledgments	xv

1	Introduction	1
1.1	A Minimalist Introduction to Molecular Evolution	1
1.2	Birth of the Combinatorics of Genome Rearrangements	4
1.3	Statement of the Problem	6
1.4	Scope of This Survey	7
1.5	Overview of the Models	7
1.6	Organization of the Book	8
I	DUPLICATION-FREE MODELS: PERMUTATIONS	11
2	Genomes as Permutations	13
2.1	The Symmetric Group	13
2.2	The Cycles of a Permutation	14
2.3	Signed Permutations	15
2.4	Distances on Permutation Groups	15
2.4.1	Rearrangements as Generators	16
2.4.2	Invariant Distances	17
2.5	Circular Permutations	18
2.5.1	Classical Circular Permutations	19
2.5.2	Genomic Circular Permutations	19
2.6	First Measures of Similarity between Permutations	20
2.6.1	Breakpoints	20
2.6.2	Common Intervals and Semipartitive Families	21
3	Distances between Unsigned Permutations	25
3.1	Transposition Distance	25
3.1.1	Lower Bounds on the Transposition Distance	26
3.1.2	Upper Bounds	29
3.1.3	Improving Bounds Using Toric Permutations	32

3.1.4	Easy Cases	33
3.1.5	Approximation Algorithms	34
3.1.6	Conjectures and Open Problems	35
3.2	Prefix Transposition Distance	36
3.2.1	Lower Bounds	37
3.2.2	Upper Bounds	38
3.2.3	Diameter	38
3.2.4	Easy Cases	39
3.2.5	Approximation Algorithms	39
3.2.6	Variant: Insertion of the Leading Element	40
3.3	Reversal Distance	40
3.3.1	Lower Bounds	40
3.3.2	Upper Bounds	43
3.3.3	Easy Cases	43
3.3.4	Computational Complexity	44
3.3.5	Approximation Algorithms	45
3.3.6	Exact Algorithms	46
3.4	Prefix Reversal Distance (Pancake-Flipping)	47
3.4.1	Lower Bounds	47
3.4.2	History	48
3.4.3	Variants	48
3.5	Variants	49
3.5.1	Block Interchange Distance	49
3.5.2	Element Interchange Distances	50
3.5.3	Weighted Reversals	52
3.5.4	Fixed-Length Reversals	54
3.5.5	Bounded Variants	54
3.5.6	Cut-and-Paste	55
3.5.7	Strip Moves	55
3.5.8	Stack-Sorting	56
3.5.9	Tandem Duplications and Random Losses	58
3.5.10	Combined Operations: Reversals and Transpositions	59
3.6	Relations between Distances on Unsigned Permutations	61
4	Distances between Signed Permutations	63
4.1	Conserved Interval Distance	63
4.2	Signed Reversal Distance	64
4.2.1	Reversals	64
4.2.2	The Distance Formula	65
4.2.3	The Scenario of Reversals	67
4.2.4	The Space of All Optimal Solutions	68
4.2.5	Experimental Results	69
4.3	Variants of Sorting by Reversals	69
4.3.1	Perfect Signed Reversal Distance	69
4.3.2	Prefix Reversals (Burned Pancakes)	70
4.3.3	Reversals That Are Symmetric around a Point	70

4.3.4	Weighted Reversals	71
4.3.5	Fixed-Length Reversals	71
4.4	Combined Operations	72
4.4.1	Reversals and Transpositions	72
4.4.2	Reversals, Transpositions, Transreversals, Revrevs	72
4.5	Double Cut-and-Joins	73
5	Rearrangements of Partial Orders	75
5.1	Genomes as Partially Ordered Sets	75
5.2	Partially Ordered Sets	75
5.2.1	Basic Definitions	75
5.2.2	Representing Posets	77
5.2.3	Topological Sorting	77
5.3	Constructing a Poset	78
5.4	Reversal Distance	79
5.5	Breakpoint Distance	80
5.5.1	Exact Algorithms	80
5.5.2	Heuristics for Computing the Breakpoint Distance	81
6	Graph-Theoretic and Linear Algebra Formulations	83
6.1	Simple Permutations and the Interleaving Graph	83
6.2	The Overlap Graph	84
6.3	The Local Complementation of a Graph	85
6.4	The Matrix Tightness Problem	85
6.5	Extension to Sorting by Transpositions	86
6.6	The Intermediate Case of Directed Local Complementation	87
II	MODELS HANDLING DUPLICATIONS: STRINGS	89
7	Generalities	91
7.1	Biological Motivations	91
7.2	Strings and Rearrangements on Strings	92
7.3	Balanced Strings	94
7.4	How to Deal with Multiple Copies?	95
8	Distances between Arbitrary Strings	97
8.1	The Match-and-Prune Model	98
8.1.1	Breakpoint Distance	100
8.1.2	Signed Reversal Distance	106
8.1.3	Adjacency Similarity	108
8.1.4	Common Intervals Similarity	111
8.1.5	Conserved Intervals Similarity	113
8.1.6	Conserved Intervals Distance	114
8.1.7	MAD and SAD Numbers	118
8.1.8	Heuristics	119

8.2	The Block Edit Model	123
8.2.1	Block Covering Distance	123
8.2.2	Symmetric Block Edit Distance	126
8.2.3	Large Block Edit Distance	129
8.2.4	String Edit Distance with Transpositions	130
8.2.5	Signed Strings	131
9	Distances between Balanced Strings	133
9.1	Minimum Common String Partition Problems	133
9.1.1	Unsigned MCSP	134
9.1.2	Signed MCSP	135
9.1.3	Reversed MCSP	137
9.1.4	Full Breakpoint Distance	138
9.2	Reversal Distance	138
9.2.1	Unsigned Reversals	138
9.2.2	Signed Reversals	141
9.2.3	Sorting by Reversals with Length-Weighted Costs	142
9.2.4	Prefix Reversals on Unsigned Strings (Pancake-Flipping)	144
9.2.5	Reversals of Length at Most 2	147
9.3	Unsigned Transpositions	147
9.3.1	Unit Cost Transpositions	147
9.3.2	Length-Weighted Transpositions	150
9.3.3	Restricted Length-Weighted Transpositions	150
9.3.4	Prefix Transpositions	152
9.3.5	Adjacent Swaps	153
9.4	Unsigned Block Interchanges	153
9.4.1	Unit-Cost Block Interchanges	153
9.4.2	Character Swaps	155
9.5	Relations between Distances	157
III	MULTICHROMOSOMAL MODELS	159
10	Paths and Cycles	161
10.1	Genomes	161
10.2	Breakpoints	162
10.3	Intervals	163
10.4	Translocation Distance	164
10.4.1	Feasibility	166
10.4.2	Unsigned Genomes	166
10.4.3	Signed Genomes	167
10.4.4	Translocations Preserving Centromeres	168
10.4.5	Variants and Special Cases	169
10.5	Double Cut-and-Joins (2-Break Rearrangement)	170
10.6	k -Break Rearrangement	171
10.7	Fusions, Fissions, Translocations, and Reversals	172
10.8	Rearrangements with Partially Ordered Chromosomes	174

11	Cycles of a Permutation	175
11.1	A Model for Multichromosomal Circular Genomes	175
11.2	A Generalization to Signed Genomes	178
11.2.1	A Different Kind of Signed Permutation	178
11.2.2	The Operations	179
11.2.3	Some Results	179
12	Set Systems and the Syntenic Distance	181
12.1	Introduction	181
12.2	Structural Properties	182
12.2.1	Compact Representation	182
12.3	Lower Bounds	184
12.4	Diameter	185
12.5	Algorithmic Results	185
12.5.1	Syntenic Distance	185
12.5.2	Easy Cases	186
12.6	Conjectures and Open Problems	189
IV	MULTIGENOMIC MODELS	191
13	Median and Halving Problems	193
13.1	Breakpoint Median	194
13.1.1	Complexity	194
13.1.2	Algorithms	195
13.2	Reversal and DCJ Median	197
13.2.1	Complexity	197
13.2.2	Algorithms	197
13.2.3	Variants	198
13.3	Duplicated Genomes	199
13.3.1	The Double Distance	199
13.3.2	Genome Halving	201
13.3.3	Solving Tetraploidy	202
13.3.4	Guided Halving	202
13.3.5	Genome Halving with Unordered Chromosomes	203
13.4	Other Variants, Generalizations, and Discussion	205
13.4.1	Other Operations	205
13.4.2	More Permutations in the Input	205
13.4.3	Medians and Centers	205
13.4.4	Discussion	206
14	Rearrangement Phylogenies	207
14.1	The Large Parsimony Problem	207
14.2	The Large Parsimony Problem with Gene Orders	209
14.2.1	Breakpoint and Reversal Phylogenies on Permutations	209
14.2.2	Variants	211

14.3	Heuristics for the Breakpoint/Reversal Phylogeny Problem	211
14.3.1	Tree Steinerization	212
14.3.2	Sequential Addition	216
14.3.3	Character Encodings	217
14.4	Variants	220

V MISCELLANEOUS 221

15 Software 223

15.1	Pairwise Rearrangements	223
15.1.1	Unichromosomal Models	223
15.1.2	Multichromosomal Models	225
15.2	Phylogeny Reconstruction and Medians	226
15.2.1	BPAAnalysis	226
15.2.2	MGR	226
15.2.3	GRIL	226
15.2.4	GRAPPA	227
15.2.5	MedRbyLS	227
15.2.6	rEvoluzer and amGRP	227
15.2.7	GENESIS	228

16 Open Problems 229

16.1	Complexity Issues	229
16.1.1	Hardness	229
16.1.2	Approximability	230
16.1.3	Polynomial Complexity	231
16.2	Diameter	231
16.3	Tightness of Bounds	232

APPENDICES 233

A Graph Theory 235

A.1	Undirected Graphs	235
A.1.1	Basic Definitions	235
A.1.2	Paths and Cycles	236
A.1.3	Connectivity	237
A.1.4	Bipartite Graphs	238
A.1.5	Trees and Forests	238
A.1.6	Matching	238
A.1.7	Adjacency Matrix	239
A.2	Directed Graphs	240
A.2.1	Basic Definitions	240
A.2.2	Paths and Cycles	241
A.2.3	Connectivity	241
A.2.4	Directed Acyclic Graphs	241

B Complexity Theory 243

- B.1 The Class **NP** 243
 - B.1.1 **NP**-Optimization Problems: From **PTAS** to **APX** 246
 - B.1.2 **NP**-Optimization Problems: Beyond **APX** 250
 - B.1.3 Parameterized Complexity 250
- B.2 Some **NP**-Complete Problems 252

- Glossary 257
- Bibliography 263
- Index 283